**Predicting Hospital Length of Stay**

**Introduction**

Healthcare costs are of primary concern in the United States (Ravi B. Parikh, 2017). Costs are consistently rising with questionable improvements in quality of care or life in relation to other nations (Joseph L. Dieleman, 2016). Due to rising costs and systemic concerns, there has been widespread debate on healthcare policy and management to optimize our current system (Barack H. Obama, 2017) (Clinton, 2016) (Henry J. Aaron, 2017). To optimize management and healthcare administration, we need a better understanding of the factors that influence overall costs. One such factor Patient length of stay (LoS) in a hospital is directly associated with cost (Paul A Taheri, 2000) (Fine MJ, 2000). There have been various attempts at predicting patient length of stay using statistical modelling (A. Azari, 2013) (Gordon H. Robinson, 1966). Good Health Corporation (GHC) has collected data on 3612 patients admitted into the hospital including Length of Stay and other possible predictors and has requested that a predictive model with Length of Stay as the outcome be created based on input predictors. To better understand the factors that affect LoS so that these can be addressed and optimized to reduce healthcare costs, we created a predictive model using multiple linear regression, based on the GHC data, that predicts patient length of stay based on relevant predictor variables

**Methods**

**Data Source**

GHC collected a total of 3682 visit records on 3612 patients who were admitted to the hospital and over the age of 17. The visit had to have occurred within 24 hours of hospital admission. Data collected for each visit included: length of stay in the hospital in days, the modified early warning score (MEWS), the Charlson Comorbidity Index rank, if the patient had an ICU visit during hospitalization, the number of ER visits in the previous 6 months, the patient’s insurance type, patient demographics, and patient vital signs.

**Data Processing and Cleaning**

The GHC dataset was processes and cleaned in order to ensure data accuracy and therefore model validity. If a patient had more than one visit, only the first visit was included for model building. Due to skew and nonnormality of the LoS outcome variable, LoS was natural logarithm transformed. For vital sign predictor variables, outliers were identified using the standard z-score method, where values outside of the middle, 99.9% of the distribution, or 3.291 standard deviations away from the mean were replaced with the mean for the predictor. This removed unrealistic values such as temperatures over 50degrees Celsius. For model building only relevant predictors were included: 30 day readmit rate, ER visits in past 6 months, Charson Index rank, MEWS, ICU visit, demographics (age, race, marital status), insurance type and vital signs (respiration rate, O2 saturation, BMI, Heartrate, Temperature, diastolic blood pressure and systolic blood pressure). We decided to omit data on religion due to lack of relevance.

**Model Selection**

Using our selected predictors, we utilized both stepwise regression selection and criterion based automatic procedures to select the best multiple linear regression model. The final model with the highest adjusted R2 value, lowest CP value and as few predictors as possible to ensure usability was selected.

**Model Diagnostics**

Once the optimal model was selected, model assumptions were checked. A residuals vs fitted value plot was created to detect for error heteroscedasticity. A quantile-quantile plot was created to detect normality of residuals. A scale-location plot was created to detect residual spread. A residuals versus leverage plot was created to help identify influential cases. Model outliers in the LoS were screened for using studentized residuals. Outliers were removed and we remodeled yet no significant difference in the model was seen and the adjusted R2 value decreases so we decided to keep outliers in the model. Leverage values in predictors were screened for as were influential predictors and none were significant enough to remove. Finally, multicollinearity was also screened for using VIF values and the findings were not significant.

**Model Validation**

We validated our final model using a Bootstrap method and calculated bias estimates for out model coefficients.

**Results:**

**Data Summary**

The mean length of stay 5.461 days with a standard deviation of 5.92. Data cleaning and missing data resulted in some loss of data for certain predictors. Summary statistics for all relevant continuous predictors used in model selection and for length of stay are included in table 1 and proportions for categorical variables are included in table 2.

---------------------------------------------------------------------

variable n mean sd minimum maximum median

----------------- ------ ------- ------- --------- --------- --------

losdays2 3612 5.461 5.92 0.04167 87.96 3.833

ageyear 3612 65.69 18.69 18 105 68

evisit 3612 1.754 1.577 0 4 1

bmi 2929 28.35 7.991 3.1 122.7 27.1

bpsystolic 3607 130.6 16.72 88.78 194 129.2

o2sat 3609 97.86 4.908 80 236.5 97.59

temperature 3610 36.73 0.899 11.85 52.27 36.73

heartrate 3607 80.07 13 37.58 242.6 79.2

respirationrate 3609 18.2 2.633 12 67.72 17.76

bpdiastolic 3611 72.52 9.798 29.56 154.4 71.85

---------------------------------------------------------------------

|  |  |
| --- | --- |
| Variable | N (%) |
| **Gender** | |
| Male | 1660 |
| Female | 1952 |
| **Race** | |
| White | 2057 |
| Black | 772 |
| Asian | 249 |
| Native American (Alaskan) | 22 |
| Native American (Hawaiian/Pacific Islander) | 4 |
| Other | 508 |
| **Insurance** | |
| Medicare | 1425 |
| Medicaid | 166 |
| Private | 1987 |
| Missing Data | 34 |
| **Marital Status** | |
| Single | 951 |
| Married | 1607 |
| Widowed | 690 |
| Divorced | 235 |
| Separated | 51 |
| Civil Union | 1 |
| Missing Data | 77 |
| 30 Day Readmit Rate | 517 |

**Final Fitted Model**

We selected a final model given by criterion based automatic model selection that included: 30 day readmit rate, ER visits, C Index, Age, Respiration Rate, Heart Rate and Systolic Blood pressure as predictors. Model coefficients and relevant values are included in table 3.

* Values associated with model
* Plot of model is here
* Comment of diagnostic plots
* Outliers results
* Comment on bootstrap validation. Model bias as found by bootstrap was quite small with the highest bias being X.

**Discussion**

* Model is minimally predictive, low R2
* Possibly better to use different modelling technique such as quadratic fit or something else
* Comment on included predictors in the final model

> sum(hos\_tidy$is30dayreadmit)

[1] 517

> sum(hos\_dummies$gender)

[1] 1660

> sum(hos\_dummies$white=='1')

[1] 2057

> sum(hos\_dummies$black=='1')

[1] 772

> sum(hos\_dummies$asian=='1')

[1] 249

> sum(hos\_dummies$natv\_amer\_alaskan=='1')

[1] 22

> sum(hos\_dummies$natv\_hawaii\_pacf\_isl=='1')

[1] 4

> sum(hos\_dummies$medicare=='1')

[1] NA

> sum(na.omit(hos\_dummies$medicare=='1'))

[1] 1425

> sum(na.omit(hos\_dummies$medicaid=='1'))

[1] 166

> sum(na.omit(hos\_dummies$private=='1'))

[1] 1987

> sum(hos\_dummies$single=='1')

[1] NA

> sum(na.omit(hos\_dummies$single=='1'))

[1] 951

> sum(na.omit(hos\_dummies$married=='1'))

[1] 1607

> sum(na.omit(hos\_dummies$widowed=='1'))

[1] 690

> sum(na.omit(hos\_dummies$divorced=='1'))

[1] 235

> sum(na.omit(hos\_dummies$separated=='1'))

[1] 51

> sum(na.omit(hos\_dummies$civil\_union=='1'))

[1] 1

> sum(is.na(hos\_dummies$married))

[1] 77

> sum(is.na(hos\_dummies$medicare))

[1] 34

3612

# References

Ravi B. Parikh, M. M. (2017). Getting Real about Health Care Costs — A Broader Approach to Cost Stewardship in Medical Education. *The New England Journal of Medicine*, 376:913-915.